EXECUTIVE BRIEF

5–7× more billable tokens from the GPUs you already own

Unlock AI for Earth, Inc.
April 2026

01

THE PROBLEM

80% of GPU capacity sits idle between inference requests. Every idle GPU second is lost revenue — you've already paid for the hardware, the power, and the data center. The industry average utilization is just 22%.

02

THE SOLUTION

The EarthServe AI Engine is a software-layer inference optimization engine. It requires no new hardware, no model retraining, and no pipeline disruption. It drops in on your existing GPU fleet and immediately begins maximizing billable tokens per GPU $-second.

03

WHY NOW

Hyperscalers have committed $660B in GPU CapEx. Hardware ceiling is real. The only remaining multiplier is software efficiency. The companies that deploy now lock in structural cost advantages that hardware spending cannot replicate.

5–7×
More billable tokens from existing GPUs — without new hardware

80%
Reduction in cost-per-token for AI inference workloads

0
New hardware, model retraining, or pipeline changes required

OUTCOME 1

More Revenue

Serve 5–7× more requests on the same infrastructure. Every additional request at near-zero marginal cost flows directly to the top line.

OUTCOME 2

Deferred CapEx

Delay your next GPU procurement cycle by 12–24 months. Redirect hundreds of millions up to hundreds of billions in planned hardware spend to higher-return initiatives.

OUTCOME 3

Margin Protection

As inference pricing compresses, margin survival requires structural cost efficiency. EarthServe AI Engine locks in a cost advantage competitors cannot easily replicate.

HOW IT WORKS

1

Drop-In Deployment
Software layer installs on existing GPU infrastructure. Connects to your Control Plane via API. No hardware changes. No downtime.

→2

Continuous Batching
Dynamic request batching fills idle GPU cycles in real time — eliminating the 78% utilization gap.

→3

Kernel Optimization
Custom CUDA kernels maximize compute per clock cycle across B200, H200, H100, A100, and equivalent hardware.

→4

Measurable Results
Billable tokens, cost-per-token, and utilization metrics visible from day one of deployment.

This brief presents representative performance benchmarks from internal testing. Results vary based on model type, GPU hardware, and workload profile. Detailed technical documentation available upon request.

Book a 30-Min Session →